Supporting Text

Moderated t - statistic

The development of the moderated t - statistic used in the manuscript is summarized here. For further details, see ref. 1. Let the normalized response for a single gene (gene subscript not indicated) from n arrays be denoted by $\boldsymbol{y}^T = (y_1, y_2, \ldots, y_n)$. Assume $E(\boldsymbol{y}) = X\boldsymbol{\alpha}$ where X is a design matrix of full column rank and $\boldsymbol{\alpha}$ is a coefficient vector. Also assume that $\text{var}(\boldsymbol{y}) = W\sigma^2$, where W is a known nonnegative definite weight matrix. Certain contrasts of the coefficients are of interest. These contrasts are defined by $\boldsymbol{\beta} = C^T\boldsymbol{\alpha}$. Assume it is of interest to test whether individual contrasts β_j are equal to zero. The linear model is fit to the responses to obtain coefficient estimators $\hat{\boldsymbol{\alpha}}$, estimators s^2 of σ^2, and estimated covariance matrices $\text{var}(\hat{\boldsymbol{\alpha}}) = Vs^2$, where V is a positive definite matrix not depending on s^2. The contrast estimators are $\boldsymbol{\beta} = C^T\hat{\boldsymbol{\alpha}}$ with estimated covariance matrices $\text{var}(\hat{\boldsymbol{\beta}}) = C^TVCs^2$. The ordinary t - statistic is given by $t_j = \hat{\beta}/\left(s\sqrt{\nu_j}\right)$ where ν_j is the j^{th} diagonal element of C^TVC.

To develop the moderated t - statistic, priors are assumed on σ^2 and β_j:

$$\frac{1}{\sigma^2} \sim \frac{1}{d_0 s_0^2}\chi^2_{d_0} \quad \text{and} \quad \beta_j|\sigma^2, \beta_j \neq 0 \sim N(0, \nu_{0j}\sigma^2).$$

Let $\tilde{s}^2$ denote the posterior mean of σ^2 given s^2. The moderated t - statistic is $\tilde{t}_j = \hat{\beta}/\left(\tilde{s}\sqrt{\nu_j}\right)$.

Equivalence Calculations

As detailed in ref. 2, a design utilizing only individuals (design I) is equivalent to a design with pools (design II) when

$$t_{s2} = t_{s1}\left[\frac{\lambda}{K(\lambda+1) - \frac{t_{a1}}{t_{a2}}}\right] \tag{1}$$

Here, $K = t_I^2/t_{II}^2$, a ratio of Student's t critical values; t_{s1} (t_{s2}) and t_{a1} (t_{a2}) denote the total number of subjects and arrays in design I (II); λ is the ratio of biological to technical variability.

This equation is quite simple, and could be extended. An optimal method for specifying equivalent designs will depend on a number of factors including the method used for finding differentially expressed genes, the desired power and error tolerance measures, distributions of biological and technical variability across genes, and the underlying model describing expression values. The method should also account for the possibility of aberrant RNA samples. In designs without pooling, if an individual sample is removed, that sample and a single array is lost. For a pooled design, all subject samples contributing to a pool are lost along with an array. The increased impact of losing an array in a pooled design, which will have a smaller sample size than the corresponding equivalent individual design, must be taken into consideration.

References

1. Smyth, G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3(1)**, Article 3.

2. Kendziorski, C.M., Zhang, Y., Lan, H. and Attie, A. (2003). *Biostatistics*, **4**, 465-477.